Educated Use of Information about Data Quality

Geoff Lee and Bill Allen
Australian Bureau of Statistics, Methodology Division
Cameron Offices, PO Box 10
Belconnen, ACT 2616, Australia
geoff.lee@abs.gov.au bill.allen@abs.gov.au

1. Introduction

This paper does not attempt to describe all the factors that contribute to the quality of data. Instead it describes only part of the work being done at the Australian Bureau of Statistics (ABS) to improve the information about quality which accompanies data products. It focuses on describing quality such that it makes those descriptions useful to external users.

2. Concepts of Quality

Over the last decade or two, there has been a steady development in concepts of quality. Earlier concepts tended to focus on accuracy, encompassing both sample and non-sample error. The focus on total survey design (e.g. Linacre and Trewin 1993) was accompanied by the gradual emergence of broader concepts of quality. Lyberg et al. (1998) trace the development of the concept of fitness for use in official statistics. As it gained prominence the relative responsibilities of producers and users of statistical data changed. As what is fit for one purpose may be unfit for another, the notion that it was sole responsibility of the producer to deliver only good quality data was replaced by a more 'caveat emptor' approach. Under this paradigm, the producer's responsibilities are first to deliver the best quality data possible within practical constraints, and then to describe the pertinent aspects of quality, so that potential consumers can make informed judgements about fitness for use.

Statistics Sweden (1994) and Elvers and Rosén (2000) present a broad level framework for describing aspects of quality that are within the direct control of a statistical office. Brackstone (1999) nominates six broad aspects of quality, and describes the generic subsystems that an effective statistical office would require in order to attain those goals. Other frameworks also include 'upstream' factors which influence quality but statistical offices cannot fully control. For example, the IMF's data quality assessment framework (Carson 2000) includes the institutional setting and funding base of statistical offices.

Some aspects of quality can be measured, but for others only indicators can be provided (Dobbs et al. 1998). Sometimes the best that can be done is to describe the methods used to generate the data, the actions that have been taken to control error, and, where relevant, the other 'upstream' factors that may impinge on data quality. While indicators and descriptions of processes do not guarantee quality, they do provide useful signals about it.

3. Existing Practices related to Quality Declarations

While existing ways of informing users about quality are a routine part of dissemination practices, it is useful to reflect on them to provide a context for the issues and opportunities that new technologies and dissemination methods present. The ABS maintains a detailed set of "Concepts, Sources and Methods" reference publications. Information and working papers, and feature articles are used to draw the attention of broader audiences to issues specific to particular data products, or changes that are being made to production processes or techniques. However, for

ad hoc users the only source of information about data quality may be a standard publication. These surround data with text that describes their content, accuracy and production processes. The layout of these publications are based on forms design principles and techniques applied to publications by the ABS in the mid-1980's. Standards were developed, beginning with theoretical ideas about 'plain English communication', which were then refined by interviewing and observing people using ABS publications.

Despite these efforts there is anecdotal evidence that the awareness, understanding and use of the quality declarations that accompany the data are poor. The more detailed quality declarations are considered hard to interpret, and irrelevant, since there is little guidance about what a consumer should do in response, other than use the data anyway.

4 Electronic Dissemination and Quality Declarations

Electronic dissemination is accepted as a strategic direction for the release of ABS data. AusStats and ABS@ both 'containers' and updated daily, are conceptually similar. They contain Acrobat ('pdf') images of all ABS paper publications, including manuals, and information papers. AusStats is available to all subscribers via the ABS website, while ABS@ operates on the intranets of major government subscribers. Some publications are accompanied by spreadsheets that contain the data from each table. From the perspective of quality, the search facilities that operate across ABS@ and AusStats represent significant improvements in the awareness and accessibility of data.

There are however some risks to the quality declarations that should accompany the data. For example data in spreadsheets may become separated from the descriptive text that give them meaning, and even mechanical issues such as linking footnotes and annotations with their corresponding data cells require attention. As electronic dissemination methods become routine, the same principles of good communication that guided the development and layout of paper publications should be applied. The usability testing of electronic products must also check that consumers' attention is directed to the aspects of quality which make data truly interpretable.

5. Making Quality Declarations Useful

The dissemination of information on quality is only worthwhile if it can be used constructively in decision-making or research. Some lessons can be drawn from ABS' efforts to improve the interpretation and use of time series data. Presentation and content were the first stages; trend data was given prominence in the graphs, tables, and commentary, and on the front page of publications. Information papers explained the rationale for this, but the most successful strategy was built around a series of seminars which explained the problem of volatility in seasonally adjusted data and offered a practical solution, the trend series. They were presented to external users and ABS staff (an important group who then acted as ambassadors for the message) repeatedly over several years. Tailored sessions were provided for policy advisors in key government departments, economic journalists and even staff in Ministers' offices. Media articles were monitored. Educating key commentators and opinion leaders was the crucial step towards convincing general users. In summary, both presentation and education components are necessary and must complement each other. Both must inform the user how to respond, not simply draw attention to the issue of quality.

In 2000, the ABS began an experimental 'Qualifying Quality' project to explore ways of improving the accessibility and interpretability of general quality declarations. The first theme of the education component is to create a common vocabulary about quality. The framework from Brackstone (1999) is applied from the point of view of a user, looking upstream through the sequence of processes that generate the data. The six dimensions are used to assess both data needs

and possible data sources, en route to a decision about the suitability of a dataset. This makes explicit the role of information about quality in assessing fitness for a decision-making or research purpose. The second theme is to show how to apply risk management principles (e.g. sensitivity analysis and contingency planning), again using the information from the quality declarations.

6. Making Quality Visible

The presentation component of the 'Qualifying Quality' project focuses on the general user who starts with the 'headline' statistics and then analyses them in more detail. Four prototype tools to access and present existing information about data quality are under development. 'Quality Issue Summaries' will cover the six major headings from the data quality framework, but not in great depth. They will draw descriptive data from the ABS' data warehouse and will be more akin to explanatory notes than detailed manuals. The 'Quality Measures' subproject restricts attention to the accuracy dimension of quality, for data from sample surveys. For each major process, simple quantitative indicators of quality have been defined and standardised across families of collections. Current prototypes generate tables and charts for comparisons across time and between surveys. The 'Data Accuracy' subproject is yet more specific, exploring ways of presenting more precise information about sampling error in electronic media. Active screen overlays showing confidence intervals, or the results of simple hypothesis tests (e.g. for significant differences from national averages) are being examined.

The final prototype draws on the others. Its goal is improved integration of data and metadata. The early prototype tool is centred on accessing a standard table electronically. Clicking on the names of row and column headings will link to more in-depth explanations. A 'quality' button at the base of the table will provide access to Quality Measures and Quality Issue Summaries, while individual cells will link to the corresponding cell level accuracy measures.

6. Conclusion

Much information about the different dimensions of the quality of official statistics is already available. Despite this, many do not appreciate the relevance of the information provided, or if they do, cannot readily interpret it, or apply it to their use of the data. Australian experience suggests that information about quality must be accompanied by practical recommendations about how to respond to it. To improve the way datasets are selected and used, an active campaign, targeting first ABS staff, and then key users and major intermediaries such as media outlets is necessary.

Frameworks for describing data quality form a good foundation for general education strategies. They provide a vocabulary that can be used to explicitly assess the fitness of statistics for a purpose. Once a dataset is judged useful, sensitivity analysis and risk assessments are practical actions that could flow from a better understanding of its strengths and weaknesses.

Better ways of presenting information about data quality are also needed, to ensure users are alerted to relevant issues, and to reinforce the messages from the education campaign. This is especially true for data that are disseminated electronically. Usability testing should ensure that the messages are getting through, not just that the IT tools are workable.

REFERENCES

Brackstone G. (1999), Managing Data Quality in a Statistical Agency, Survey Methodology, Vol. 25, no. 2, pp. 129-149

Carson C. (2000), Toward a Framework for Assessing Data Quality, The Proceedings of Statistical Quality Seminar, Jeju Korea, Korea National Statistical Office and IMF

Dobbs J., Gibbins C., Martin J., Davies P. and Dodd T. (1998), Reporting on Data Quality and Process Quality, American Statistical Association, Survey Research Methods Proceedings pp. 32-40

Elvers E. and Rosén B. (2000), Quality Concept for Official Statistics, Encyclopaedia of Statistical Sciences, Update Volume 3, pp. 621-629

Linacre S., and Trewin D. (1993), Total Survey Design - Application to a Collection of the Construction Industry, Journal of Official Statistics, Vol. 9, No. 3, pp. 611-621

Lyberg L., Japec L. and Biemer P, (1998), Quality Improvements in Surveys - A Process Perspective, American Statistical Association, Survey Research Methods Proceedings, pp. 23-31

Statistics Sweden (1994), (Quality Definition and Recommendations for Quality Declarations of Official Statistics, in Swedish), Statistics Sweden, Stockholm.

RESUME

Cet article considère un aspect de la qualité des données, à savoir fournir des informations sur la qualité des données aux utilisateurs. Bien que beaucoup d'informations sur la qualité des données soient déjà disponibles, beaucoup de personnes n'apprécient pas leur pertinence, ou ne peuvent pas l'interpréter, ou bien ne l'utilisent pas dans la pratique. Une meilleure formation des utilisateurs ainsi qu'une meilleure présentation de l'information concernant la qualité sont nécessaires pour changer cette situation. En particulier, les informations sur la qualité doivent être accompagnées de recommandations pratiques sur la façon de répondre.

Les cadres conceptuels décrivant la qualité des données fournissent une bonne base pour mettre en application des stratégies de formation générale. Ils fournissent un vocabulaire qui pourrait être enseigné aux utilisateurs de données statistiques, pour les aider à évaluer l'adéquation des statistiques à leurs objectifs. Quand un ensemble de données a été jugé utile, une analyse de sensibilité, une évaluation des risques et un plan de secours sont des actions pratiques qui pourraient découler d'une meilleure compréhension des forces et des faiblesses d'un ensemble de données.

De meilleures méthodes de présentation des informations sur la qualité de données sont nécessaires également, pour assurer que les utilisateurs soient tenus informés des questions pertinentes, et pour renforcer les messages de la campagne de formation. Ceci est vrai particulièrement pour les données qui sont diffusées électroniquement. Heureusement, les avancées technologiques dans la diffusion électronique donnent également des opportunités de développer de meilleures méthodes de présentation. De telles méthodes devraient lier les données et l'information décrivant les dimensions diverses